Cassandra Crossing/ Archivismi: il giorno dopo l'upload

(561)—Ieri abbiamo fatto il nostro primo upload e ne abbiamo visto i risultati. Ma oggi è cambiato qualcosa?

Cassandra Crossing/ Archivismi: il giorno dopo l'upload
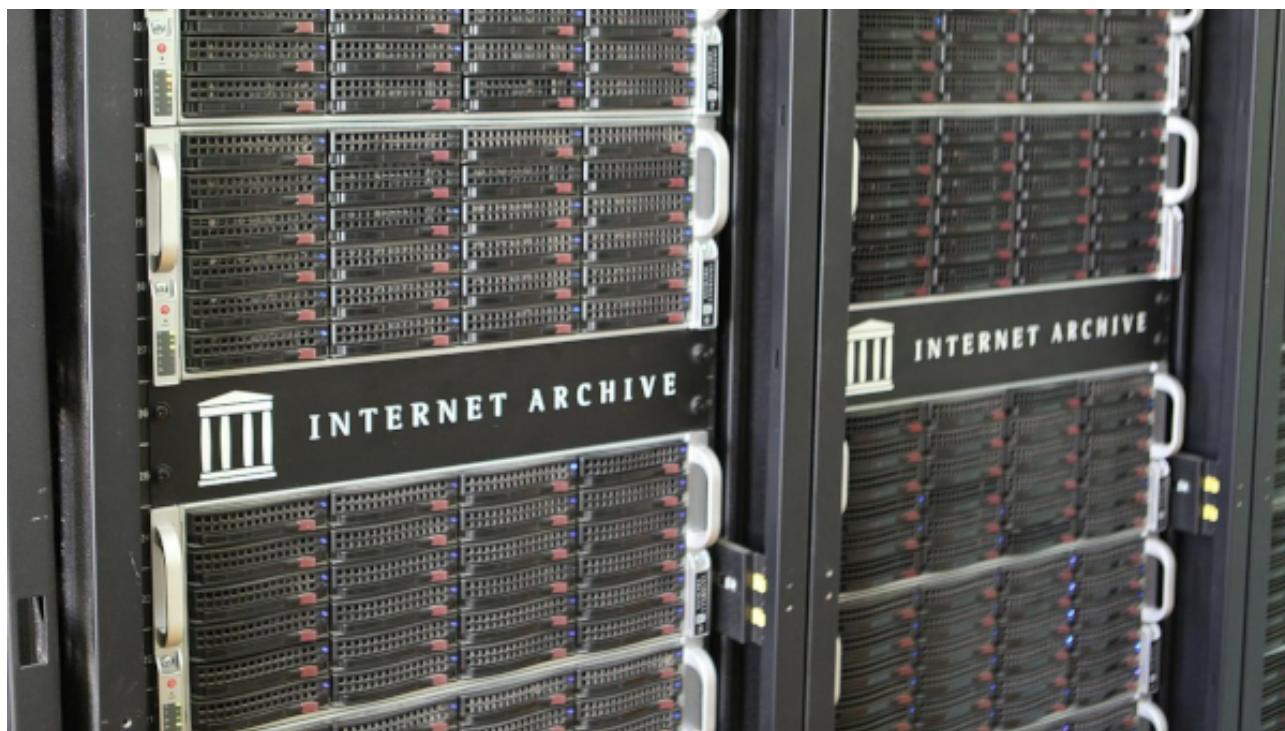
Figure 1:

(561)—Ieri abbiamo fatto il nostro primo upload e ne abbiamo visto i risultati. Ma oggi è cambiato qualcosa?

27 dicembre 2023—Nella scorsa puntata Cassandra ha cercato di raccontarvi una parte del funzionamento di Internet Archive. Abbiamo appena scalfito la superficie delle sue caratteristiche, e per non annoiarci abbiamo provato ad archiviare il file .pdf di un articolo di Cassandra, ed a descrivere cosa succedeva.

Ci siamo così resi conto di aver avviato un processo tanto complesso quanto lento, ma per fortuna completamente automatico. Tanto lento che dopo più di mezz'ora non si era ancora concluso. Tornando oggi sulla pagina del documento, troviamo il *browser di oggetti* di Internet Archive attivo, ed il processo che si è completato.
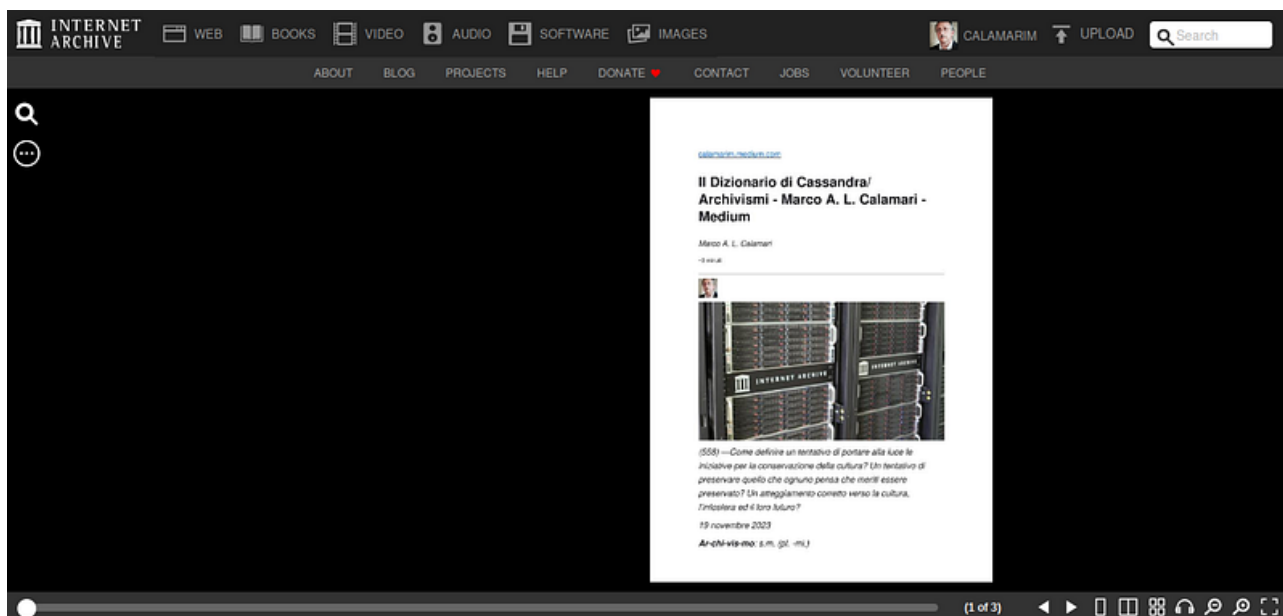
Figure 2:

E' possibile sfogliare rapidamente le pagine, farle leggere ad una voce molto robotica, e selezionare parti di testo su qualsiasi pagina. Sembrano cose da poco, considerando che la sorgente era un pdf "moderno", ottenuto direttamente da un documento Libreoffice, ma in effetti l'apparentemente "semplice" pdf è stato scomposto in una quantità di file, alcuni dei quali non avevamo ancora analizzato.
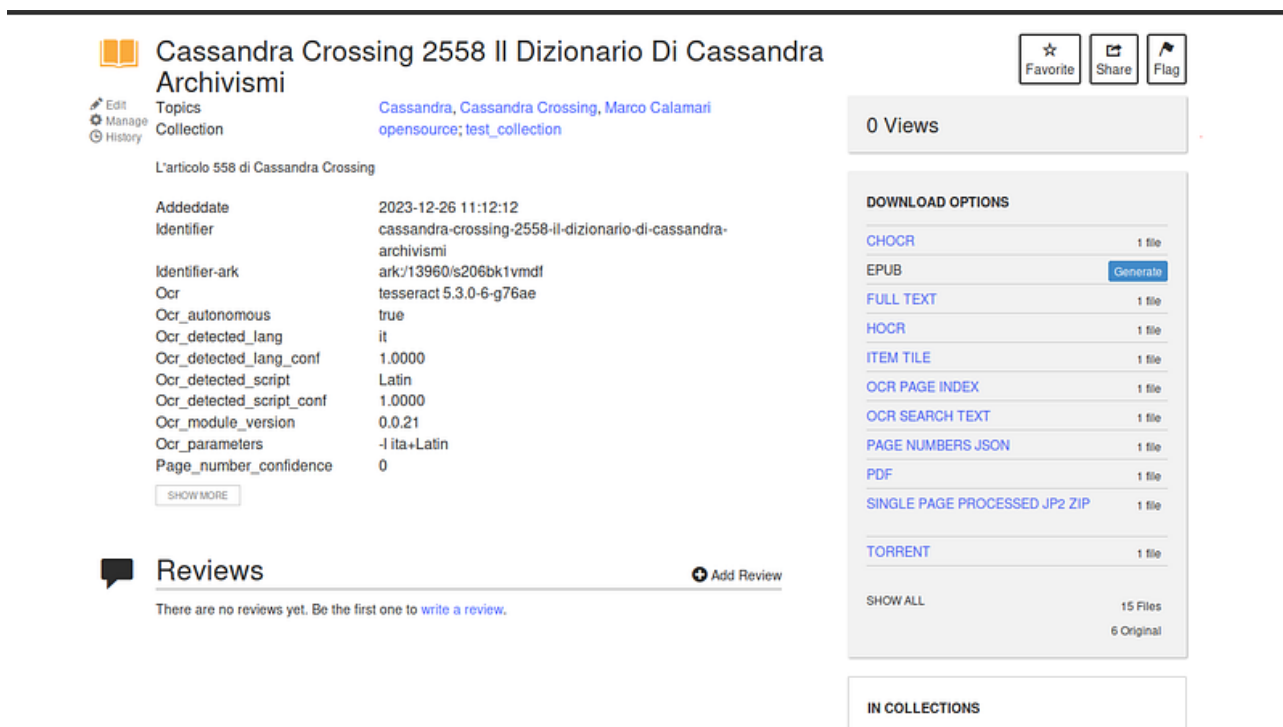
Figure 3:

Anche solo dai nomi, possiamo facilmente capire che un qualche processo OCR di riconoscimento

dei caratteri è stato eseguito automaticamente. Questi file, alcuni dei quali vengono usati dal *browser di oggetti* di Internet Archive, permettono a quest'ultimo di visualizzare il documento.

A questo punto qualcuno degli informatissimi 24 lettori sbotterà "*Ma tutto questo è assolutamente banale, lo si poteva fare anche con Acrobat Reader, senza tutto questo ambaradan.*" Il caro lettore ha ragione sul fatto specifico, ma torto sulla questione più generale. Si, perché archiviando il pdf moderno di 3 pagine abbiamo in realtà usato un cannone per ammazzare una zanzara, perdipiù gracilina e malata.

Ora è arrivato il momento di provare a scatenare tutta la potenza archiviativa di *Internet Archive*. Per questo Cassandra ha sfruttato un lavoro di archiviazione che attendeva il suo alter-ego Marco Calamari. Si trattava di archiviare un centinaio di numeri di una piccola rivista, uscita negli ultimi 30 anni ed esclusivamente in formato cartaceo.

Erano già stati raccolti i file .pdf generati dai vari programmi di impaginazione elettronica usati per realizzare la rivista, e per fortuna conservati come sottoprodotto. Erano state anche realizzate, artigianalmente ed in vari modi, le scansioni dei primi numeri cartacei, anche questi in formato pdf, ma ovviamente non ricercabili, essendo le pagine delle "*fotografie*".

Tutto questo materiale, anche se già in formato digitale, avrebbe richiesto un tempo lunghissimo per essere messo insieme, allineato e pubblicato in un formato ricercabile e riutilizzabile, particolarmente in ambiti di archiviazione "seria".

Infatti il vero, grosso problema non era quello di creare una collezione di file pdf, ma quella di archiviarla in maniera utile, ricercabile e consultabile. Altrimenti, come spesso accade, questi file, pur faticosamente raccolti, sarebbero prima o poi finiti dimentcati in una chiavetta in fondo ad un cassetto, od in un angolo di cloud commerciale, effimero e dove nessuno (tranne i GAFAM) li avrebbe potuti trovare ed utilizzare.

Ma è bastato mettere insieme i 75 file di vario formato e contenuto in un unico pdf, usando l'utilissimo software libero Pdftk, realizzando così un pdf unico di quasi 1 terabyte, ed uploadare quest'ultimo su Internet Archive, esattamente come avevamo fatto per l'articoletto di 3 pagine. Anche questo file è stato preso in carico dal sistema e "tritato" per tutta la notte; stamani era già disponibile.

Tutte le anomalie e le differenze erano state risolte automaticamente, ed un documento di 662 pagine, contenente l'intera raccolta della rivista, era disponibile, rapidamente sfogliabile, selezionabile, ricercabile e ascoltabile, ed era stato creato con un impegno di pochi minuti di tempo.

Figure 4:

Se aggiungiamo a questo il fatto che il documento è stato archiviato in maniera ridondante in più datacenter, e si trova in una in una biblioteca digitale che lo mette a disposizione di chiunque, liberamente ricercabile e visualizzabile, la cosa diventa quasi stupefacente, anche senza aggiungere che è disponibile pure in formato ebook (.epub) e che se necessario può essere ulteriormente "lavorato" per altri scopi.

Giusto per descrivere in linea di massima cosa è stato prodotto durante l'archiviazione, il pdf originale è stato diviso in pagine, prima di tutto per velocizzarne la visualizzazione. Ciascuna pagina è costituita da un file pdf in un formato particolare, una immagine di sfondo, la scansione della pagina originale, più un layer di testo selezionabile, sovrapposto alla pagina e generato sottoponendo ad OCR la scansione stessa.

La cosa veramente notevole è che il sistema è stato in grado di gestire correttamente un misto di file pdf con differenti strutture interne, da semplici scansioni a pdf strutturati, e di riportarli tutti ad un minimo comune multiplo costituito dai pdf a strati delle singole pagine.

Beh, se tutto questo vi sembrasse poco, è perché questa serie di articoli non è adatta a voi; è invece adatta ai futuri *bibliotecari digitali* che, per caso o per fortuna, siano capitati su queste paginette. Ma potreste ancora cambiare idea.

Stay tuned per la prossima puntata di "*Archivismi*".

Scrivere a Cassandra—Twitter—Mastodon
Videorubrica "Quattro chiacchiere con Cassandra"
Lo Slog (Static Blog) di Cassandra
L'archivio di Cassandra: scuola, formazione e pensiero

By Marco A. L. Calamari on December 27, 2023.

Canonical link

Exported from Medium on August 27, 2025.